Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons

Supplementary Data

Sascha Steinbiss, Ute Willhoeft, Gordon Gremme and Stefan Kurtz

August 27, 2009

Drosophila melanogaster sequence data

The genomic sequence for *Drosophila melanogaster*, release 5.8 was downloaded in June 2008 from the FlyBase FTP server at `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.8_FB2008_05`.

Mus musculus sequence data

The genomic sequence for chromosome 4 of the *Mus musculus* genome (build 37, version 1) was downloaded in May 2009 from the NCBI FTP server at `ftp://ftp.ncbi.nih.gov/genomes/M_musculus/Assembled_chromosomes/mm_ref_chr4.fa.gz`.

LTR retrotransposon reference sequences

For sequence-based comparisons of predicted LTR retrotransposon candidate sequences with a reference set of canonical *Drosophila melanogaster* LTR retrotransposon sequences, a file in EMBL format was obtained from the FlyBase FTP server at `ftp://ftp.flybase.net/releases/current/precomputed_files/transposons/transposon_sequence_set.embl.txt.gz`. It contains one representative canonical sequence per transposon for which exemplary sequences are stored in FlyBase. Version v9.41 of the file, dated April 25, 2005, was used. This file in EMBL format was converted to multiple FASTA format. Besides 111 other transposable elements, 61 LTR retrotransposon sequences from *D. melanogaster* are present in this file. 49 of these 61 are marked as 'complete', denoting representative sequences deemed full-length by the curators. These 49 sequences were extracted to a separate file for performance evaluation of the classification step.

PPT scoring

Let c be a LTR retrotransposon candidate with its inner 3' LTR boundary located at position b_3. Let (i, j) be the start and end positions of a PPT sequence in c. As the PPT was predicted in the radius r around b_3, it is guaranteed that $b_3 - r < j < b_3 + r$. Each PPT candidate is scored based on the distance of its end position j to the inner 3' LTR boundary as follows:

$$score_{ppt}(j) := \frac{r^2 - |b_3 - j|^2}{r^2}$$

PBS scoring

Let A be a local alignment of length h between a tRNA sequence t with length $|t|$ and a substring of length $2r+1$ around the 5' LTR boundary e_5 of a LTR retrotransposon candidate. Let o_t be the offset of the start of the aligned substring t' on the tRNA from the tRNA 3' end and o_s be the offset of the start of the aligned substring s' on the candidate sequence from the 5' LTR inner boundary. Furthermore, let d be the unit edit distance of t' and s'. A is then scored as follows:

$$score_{pbs}(A) = h \cdot \frac{|t| - o_t}{\max(1, d \cdot o_s) \cdot |t|}$$

Clustering parameters

This section describes the parameters used with the *dbcluster* tool from Vmatch during the separate feature clustering step.

LTR sequences

Clusters of LTR sequences (from b_3 to e_3 and b_5 to e_5 respectively, see Figure 1) are formed when matches between two sequences are found that span at least 80 % of the shorter sequence with sequence identity of at least 80 %. See below for more details. As further options for the matching algorithm, a seed length of 10 characters and an X-drop value of 7 was used.

PBS/PPT sequences

PPT and PBS sequences are short signals (mostly shorter than 25 bp) which must be handled in a stricter way to avoid unspecific matches. Here matches must span at least 90 % of both sequences with 90 % sequence identity to join the matching sequences into a cluster. Additionally, the seed length was lowered to 3 characters and the X-drop value to 2. As a minimal match length, 90 % of the smaller of both matches was used.

Protein domain sequences

Since the protein domain hit lengths tend to vary in wide ranges, it must be possible to consider matches between short and long sequences for clustering. This is achieved by putting two sequences into the same cluster if the shorter sequence involved in the match spans at least

$$\left\lfloor \frac{l_{\min}}{l_{\max}} \cdot 80 \right\rfloor + 1$$

percent of the longer sequence, where $l_{\min}$ denotes the length of the shorter sequence and $l_{\max}$ the length of the longer matching sequence. The factor 80 accounts for the fact that the match on the shorter sequence must span at least 80 % of the sequence. Additionally, the seed length was 10 and the X-drop value was 3.

Selection of putative good representatives per candidate group

Let C be the set of LTR retrotransposon candidate sequences. Let D be the set of protein domain models. Then $F = \{ltr_5, ltr_3, pbs, ppt\} \cup D$ is the set of possible features assigned by *LTRdigest*. This assignment is represented by a function φ such that $\varphi(c, f) = (i, j)$ if candidate c has feature f in its substring $c[i..j]$. If c does not have feature f, we write $\varphi(c, f) = \perp$, where $\perp$ stands for undefined.

Let $domcnt(c) = |\{f \in F \mid \varphi(c, f) \neq \perp\}|$ be the number of internal features detected in candidate c and $grpcnt(k, G) = |\{c \in G \mid domcnt(c) = k\}|$ be the number of candidates in group G with exactly k features. Furthermore, let $len(c)$ be the candidate length of candidate c and $len_{3'}(c)$ be the 3' LTR length of c. We also define $mlen(G)$ as the median candidate length in group G and $mlen_{3'}(G)$ as the median 3' LTR length in group G. The candidate $c \in G$ of length $len(c)$ with 3' LTR length $len_{3'}(c)$ is chosen as a representative for group G if the following conditions are satisfied:

- the number of detected features in c is equal to the one observed most frequently in its group, i.e. $grpcnt(domcnt(c), G) \geq grpcnt(domcnt(c'), G)$ for all $c' \in G$.
- the overall length of c does not deviate from the median length in the group G by more than a given threshold th_c, i.e. $|len(c) - mlen(G)| \leq th_c$.
- the 3' LTR length of c does not deviate from the median length in the group G by more than a given threshold th_l, i.e. $|len_{3'}(c) - mlen_{3'}(G)| \leq th_l$.

In the *Drosophila melanogaster* use case, the parameters $th_c = 300$ and $th_l = 50$ were used. To address boundary cases, if a group consists of only two candidates or only one representative was selected, all members of this group are considered for further analysis.

Table A1

LTRharvest settings used in the analysis of the *Drosophila melanogaster* and *Mus musculus* sequences. The *D. melanogaster* parameter set is identical to the one published by Ellinghaus et al. [1] except that no overlaps are allowed in our case.

Parameter	*D. melanogaster*	*M. musculus*
LTR length	116–800 bp	100–1000 bp
LTR distance	2280–8773 bp	1000-15000 bp
LTR similarity	91 %	80 %
TSD length	4–20 bp	4–20 bp
motif/TSD search radius	60 bp	60 bp
Seed length	76 bp	30 bp
X-drop X value	7	5
X-drop $score_{mat}$	2	2
X-drop $score_{mis}$	-2	-2
X-drop $score_{ins}$	-3	-3
X-drop $score_{del}$	-3	-3
overlaps	no	no

Table A2

LTRdigest settings used in the analysis of the *Drosophila melanogaster* and *Mus musculus* candidates. HMM emission probabilities p_R and p_T were manually derived from a small set of example sequences [2] of retroviral and retrotransposon PPTs. Otherwise, a uniform background distribution with an emission probability of 0.25 for each nucleotide was used.

Parameter	*D. melanogaster*	*M. musculus*
PPT search radius	30 bp	
PPT length range	8–30 bp	
PPT A/G emission probability p_R	0.97	
U-box length range	3–30 bp	
U-box U/T emission probability p_T	0.91	
PBS search radius	30 bp	
PBS alignment length	11–30 bp	
PBS offset range	0–5 bp	
tRNA offset range	0–40 bp	0–5 bp
PBS/3' tRNA end max. edit distance	1	
tRNA library	100 tRNAs	248 tRNAs
Smith-Waterman $score_{mat}$	5	
Smith-Waterman $score_{mis}$	-10	
Smith-Waterman $score_{ins}$	-20	
Smith-Waterman $score_{del}$	-20	
Domain pHMM set	22 domains	21 domains
pHMM hit E-value cutoff	10^{-6}	
max. gap length between chained fragments	50 bp	

Table B1

Pfam protein domain models used in the set of LTR retrotransposon/retrovirus-specific domains for the *D. melanogaster* analysis. Profile HMM files were selected from Pfam [3] (`http://pfam.sanger.ac.uk/`) using the search terms "retrotransposon", "env transposon", "reverse transcriptase", "retroelements" and "gag transposon", resulting in a first set of potentially relevant protein domain models. Also, we used the "Domain Organisation" section in each Pfam entry to identify other domains commonly present in the neighborhood of these domains in other retrotransposon sequences. These were added to the final set if they were relevant. When using *LTRdigest*, such a set is most conveniently organised as a directory containing the downloaded pHMM files. Then a set can easily be used in a *LTRdigest* run by giving `/path/to/set/directory/*.hmm` as an argument to the option `-hmms`.

Pfam accession#	Pfam ID	Description
PF03732	Retrotrans_gag	Retrotransposon gag protein
PF07253	gypsy	Gypsy protein
PF00077	RVP	Retroviral aspartyl protease
PF08284	RVP_2	Retroviral aspartyl protease
PF00078	RVT_1	Reverse transcriptase
PF07727	RVT_2	Reverse transcriptase
PF06817	RVT_thumb	Reverse transcriptase thumb domain
PF06815	RVT_connect	Reverse transcriptase connection domain
PF00075	Rnase H	Ribonuclease H domain
PF00552	Integrase	Integrase DNA binding domain
PF02022	Integrase_Zn	Integrase Zinc binding domain
PF00665	rve	Integrase core domain
PF00098	zf-CCHC	Zinc knuckle domain
PF00385	Chromo	'chromo' (CHRomatin Organisation MOdifier) domain
PF01393	Chromo_shadow	Chromo shadow domain
PF00692	dUTPase	dUTPase domain
PF01021	TYA	TYA transposon protein
PF03078	ATHILA	ATHILA ORF-1 family
PF04094	DUF390	Protein of unknown function (DUF390)
PF08330	DUF1723	Protein of unknown function (DUF1723)
PF04195	Transposase_28	Putative gypsy type transposon
PF05380	Peptidase_A17	Pao retrotransposon peptidase

Table B2

Pfam protein domain models used in the *Mus musculus* analysis.

Pfam accession#	Pfam ID	Description
PF01140	Gag_MA	Matrix protein (MA), p15
PF02337	Gag_p10	Retroviral GAG p10 protein
PF01141	Gag_p12	Gag polyprotein, inner coat protein p12
PF00607	Gag_p24	gag gene protein p24 (core nucleocapsid protein)
PF02093	Gag_p30	Gag P30 core shell protein
PF00692	dUTPase	dUTPase domain
PF00077	RVP	Retroviral aspartyl protease
PF00078	RVT_1	Reverse transcriptase
PF06817	RVT_thumb	Reverse transcriptase thumb domain
PF00552	Integrase	Integrase DNA binding domain
PF02022	Integrase_Zn	Integrase Zinc binding domain
PF00665	rve	Integrase core domain
PF00075	Rnase H	Ribonuclease H domain
PF00429	TLV_coat	ENV polyprotein (coat polyprotein)
PF08791	Viral_env	Viral envelope protein
PF09590	Env-gp36	Env-gp36 lentivirus glycoprotein
PF03408	Foamy_virus_ENV	Foamy virus envelope protein
PF00516	GP120	Envelope glycoprotein GP120
PF03056	GP36	Env gp36 protein (HERV/MMTV type)
PF00517	GP41	Envelope Polyprotein GP41
PF00098	zf-CCHC	Zinc knuckle domain

Table C

LTR retrotransposon candidate groups from *D. melanogaster* showing a global identity of less than 80 % to reference sequences (21 groups out of 62).

Candidate groups		Reference		
Group	# members	Identity (%)	Family	Comment
Dmel-18	2	72.9	invader6	one nested, one with internal deletions
Dmel-22	5	–	–	shows PR–RT–INT domain matches, short matches to gypsy-like LTR retrotransposons (longest $\approx$ 400 bp, 68 % similarity, to *gypsy7* from *Anopheles gambiae*), has *gypsy env* pHMM hit, has $tRNA^{Lys}$ PBS, has 13 bp PPT
Dmel-25	2	–	–	conglomerate of *PROTOP* DNA transposons with *DMCR1A* non-LTR retrotransposons
Dmel-28	3	–	–	fragmented *copia*-like sequences with *G6* non-LTR retrotransposon insertion
Dmel-29	2	65.3	Circe	internal deletions
Dmel-31	3	79.7	gypsy12	other transposon insertions
Dmel-35	3	–	–	fragmented *Dsim/ninja* matches
Dmel-39	3	35.4	invader1	*gypsy*, *gypsy6* and *invader1* LTR retrotransposon sequences with *FW_DM* non-LTR retrotransposon insertions
Dmel-41	2	–	–	conglomerate of DNA transposons and *baggins* elements
Dmel-42	3	–	–	best match: 1800 bp, 64.5 % similarity to *gypsy35* from *Anopheles gambiae*
Dmel-46	2	–	–	conglomerate of DNA transposons and *DMRT1B* non-LTR retrotransposons
Dmel-47	2	74.8	Dsim/ninja	internal deletions
Dmel-48	2	–	–	conglomerate of DNA transposons with *DMCR1A* and G5_DM non-LTR retrotransposons
Dmel-50	2	–	–	conglomerate of *gypsy-like* fragments and *DMCR1A* non-LTR retrotransposons
Dmel-51	2	74.9	gypsy10	with partial *gtwin* insertion
Dmel-53	2	39.8	invader2	fragmented *DMRT1C* non-LTR retrotransposon sequences with various DNA transposon, non-LTR retrotransposon and LTR retrotransposon fragments
Dmel-55	2	–	–	conglomerate of DOC3_DM, BS2_DM and *DMRT1B* non-LTR retrotransposons
Dmel-56	2	–	–	fragmented *gypsy12* sequence with *DOC* non-LTR retrotransposon insertion
Dmel-57	3	52.9	micropia	internal deletions
Dmel-58	2	67.5	roo	internal inversion
Dmel-61	2	60.7	rover	two nested elements reported as one

Table D

Families whose LTR retrotransposon reference sequences could not be recognised in the candidate groups.

Family	Reason for not being present in the groups
1731, *accord*, *Dm88*, *gypsy2*, *gypsy3*, *McClintock*, *Stalker*, *rooA*	full-length matches only in candidates discarded as ambiguously matching or singlets
Circe, *invader5*	no full-length hit found in *LTRharvest* output
gtwin, *gypsy5*	only one match, discarded as only member of its group

Table E1

Most frequent protein domain hit patterns in the *D. melanogaster* candidates. Patterns with less than five occurrences are not shown. See Tables B1 and B2 for details about the Pfam IDs we are referring to. Abbreviations: PR = Protease, RT = reverse transcriptase, INT = integrase.

Domain order pattern	Domain hits in reading direction	# occ.
PR–RT–INT	RVP → RVT_1 → rve	207
PR–INT	Peptidase_A17 → rve	115
INT–RT	zf-CCHC → rve → RVT_2	41
RT–INT	RVT_1 → rve	38
RT–INT	RVT_1 → rve → Gypsy	32
RT–INT	zf-CCHC → RVT_1 → rve	26
RT	RVT_1	25
PR–RT–INT	RVP → RVT_1 → rve → Gypsy	25
PR–RT–INT	zf-CCHC → RVP → RVT_1 → rve	18
INT	rve	13
PR–RT	RVP → RVT_1	11
INT	zf-CCHC → rve	10
RT–PR–INT	zf-CCHC → RVT_1 → Peptidase_A17 → rve	10
RT	zf-CCHC → RVT_1	8
PR	Peptidase_A17	7

Table E2

Most frequent protein domain hit patterns in the *M. musculus* candidates. Patterns with less than five occurrences are not shown. See Tables B1 and B2 for details about the Pfam IDs we are referring to. Abbreviations: PR = Protease, RT = reverse transcriptase, INT = integrase.

Domain order pattern	Domain hits in reading direction	# occ.
RT	RVT_1	164
PR–RT–INT	Gag_p24 → zf-CCHC → dUTPase → RVP → RVT_1 → RVT_thumb → RnaseH → Integrase_Zn → rve → Integrase	46
RT–INT	RVT_1 → RnaseH → rve → dUTPase	33
RT–INT	RVT_1 → RVT_thumb → RnaseH → Integrase_Zn → rve → Integrase	18
INT	Integrase	17
PR–RT–INT	Gag_MA → Gag_p30 → zf-CCHC → RVP → RVT_1 → RnaseH → rve → TLV_coat	14
–	Gag_p24	12
PR	Gag_p24 → zf-CCHC → dUTPase → RVP	11
PR–RT–INT	Gag_p10 → Gag_p24 → zf-CCHC → dUTPase → RVP → RVT_1 → RVT_thumb → RnaseH → Integrase_Zn → rve → Integrase	11
PR–RT–INT	RVP → RVT_1 → RVT_thumb → RnaseH → Integrase_Zn → rve → Integrase	9
INT	rve	9
RT–INT	dUTPase → RVT_thumb → RnaseH → Integrase_Zn → rve → Integrase	7
INT	Gag_p24 → dUTPase → rve → Integrase	6

Table F1

tRNAs complementary to PBS motifs identified in *Drosophila melanogaster* LTR retrotransposon candidates.

amino acid	anticodon	# occurrences
Ser	AGA	81
Arg	TCG	55
Met	CAT	40
Lys	TTT	35
Leu	CAA	32
Tyr	GTA	16
Lys	CTT	8
Ile	AAT	7
Glu	CTC	2
Pro	CGG	2
Gly	TCC	2
Leu	CAG	2
Arg	TCT	2
Cys	GCA	1
Trp	CCA	1
Glu	TTC	1
Val	TAC	1
Thr	AGT	1
Ala	AGC	1
Pro	AGG	1
Thr	TGT	1

Table F2

tRNAs complementary to PBS motifs identified in the putative *Mus musculus* (near) full-length LTR retrotransposon candidates.

amino acid	anticodon	# occurrences
Ser	GGA	13
Phe	GAA	9
Gln	TTG	6
Pro	TGG	4
Gly	GCC	3
Gln	CTG	2
Lys	TTT	1
Leu	TAA	1

Table G

Results from filtering *D. melanogaster* release 3 LTR retrotransposon candidates using *LTRdigest*-based annotations. Candidate positions were compared with the 'full-length' reference annotation from Ellinghaus et al. [1]. Filtering by protein domain presence leads to a high specificity improvement with no sensitivity loss.

	no filtering	filtered by protein domain presence
# filtered candidates	506	400
TP	299	299
FP	207	101
FN	5	5
sensitivity	98%	98%
specificity	59%	75%

Abbreviations: TP = true positives, FP = false positives, FN = false negatives. Prediction quality was calculated by comparing the start and end positions of each element to a reference database of 304 known full-length elements [4]. Candidates whose start or end positions on a specific chromosome arm match those of an element in the reference database within a difference of 20 base pairs are considered true positives.

References

[1] Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008) *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics,* **9**, 18.

[2] Wilhelm, M. and Wilhelm, F. X. (2001) Reverse transcription of retroviruses and LTR retrotransposons. *Cell. Mol. Life. Sci.,* **58**, 1246–1262.

[3] Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res.,* **36**, D281–D288.

[4] Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.,* **3**, RESEARCH0084.